

APPLICATION  
FOR  
UNITED STATES LETTERS PATENT

TITLE: PROTEIN MODELING TOOLS  
APPLICANT: DR. JEFFREY SKOLNICK AND ANDRZEJ KOLINSKI

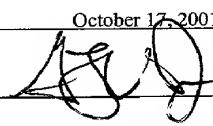
CERTIFICATE OF MAILING BY EXPRESS MAIL

Express Mail Label No. EI584812234US

I hereby certify under 37 CFR §1.10 that this correspondence is being deposited with the United States Postal Service as Express Mail Post Office to Addressee with sufficient postage on the date indicated below and is addressed to the Commissioner for Patents, Washington, D.C. 20231.

Date of Deposit October 17, 2001

Signature



Gildardo Vargas

Typed or Printed Name of Person Signing Certificate

0566243-10101  
10/17/01 09:23:50

## PROTEIN MODELING TOOLS

### GOVERNMENT INTERESTS

The instant invention was partially supported by a grant from the United States government under grant No GM-48835 awarded by the National Institutes of Health. As a result, the government may have certain rights in the invention.

### RELATED APPLICATIONS

This application claims the benefit of priority under 35 U.S.C. § 119(e) of U.S. provisional patent application serial numbers 60/117, 570, filed January 27, 1999, and 60/118,844, filed February 5, 1999. Each of the aforementioned applications is explicitly incorporated by reference in their entirety and for all purposes.

### FIELD OF THE INVENTION

This invention concerns tools useful for modeling the three-dimensional structure of proteins. Specifically, the invention concerns algorithms, computer systems, and methods for determining, predicting, and/or refining three-dimensional structures of proteins.

### BACKGROUND OF THE INVENTION

The following description of the background of the invention is provided to aid in understanding the invention. It is not an admission that any of the information provided herein is prior art to the presently claimed invention, nor that any of the publications specifically or implicitly referenced are prior art to that invention.

A central tenet of modern biology is that heritable genetic information resides in a nucleic acid genome, and that the information embodied in such nucleic acids directs cell function. This occurs through the expression of various genes in

the genome of an organism and regulation of the expression of such genes. The  
5 pattern of which subset of genes in an organism is expressed at a particular time in a  
particular cell defines the phenotype, and ultimately cell and tissue types. While the  
least genetically complex organisms, *i.e.*, viruses, contain on the order of 10-50  
genes and require components supplied by a cell of another organism in order to  
reproduce, the genomes of independent, living organisms (*i.e.*, those having a  
10 genome that encodes for all the information required for the organism to survive and  
reproduce) that are the least genetically complex have more than 400 genes (for  
example, *Mycoplasma genitalium*). More complex, multicellular organisms (*e.g.*,  
mice or humans) contain genomes believed to be comprised of tens of thousands or  
more genes, each of which codes for one or more different expression products.

15 Some of these genes are transcribed, but not translated; thus, the final gene  
products of these genes are RNA molecules (for example, ribosomal RNAs, small  
nuclear RNAs, transfer RNAs, and ribozymes (*i.e.*, RNA molecules having  
endoribonuclease catalytic activity)). However, most RNAs are mRNAs, and these  
are translated into proteins. The particular sequence of the ribonucleotides  
20 incorporated into an RNA as it is synthesized is dictated by the gene found in the  
genomic DNA from which it was transcribed. In the translation of an mRNA, the  
particular nucleotide sequence determines the particular amino acid sequence of the  
protein translated therefrom, and it is a protein's amino acid sequence that ultimately  
determines its three-dimensional structure, taking into account the thermodynamics  
25 of the system in which the protein is assembled. Significantly, three-dimensional  
structure dictates the particular biological function(s) of any biomolecule, including  
proteins.

The elegant simplicity of the foregoing schema is obscured by the  
complexity and size of the genomes found in living systems. For example, the  
30 haploid human genome comprises about  $3 \times 10^9$  (three billion) nucleotides spread  
across 23 chromosomes. However, it is currently estimated that less than 5% of this  
encodes the approximately 80,000-100,000 different protein-coding genes believed